Content-based inference of hierarchical structural grammar for recurrent TV programs using multiple sequence alignment

Bingqing Qu, Félicien Vallet, Jean Carrive, Guillaume Gravier

CONTENT-BASED INFERENCE OF HIERARCHICAL STRUCTURAL GRAMMAR FOR RECURRENT TV PROGRAMS USING MULTIPLE SEQUENCE ALIGNMENT

Bingqing Qu[1,2]*, Félicien Vallet*[2]*, Jean Carrive*[2]*, Guillaume Gravier*[1,3]

[1]University of Rennes 1
[2]French National Audiovisual Institute
[3]CNRS & IRISA & INRIA Rennes
{bqu, fvallet, jcarrive}@ina.fr, guillaume.gravier@irisa.fr

ABSTRACT

Recently, unsupervised approaches were introduced to analyze the structure of TV programs, relying on the discovery of repeated elements within a program or across multiple episodes of the same program. These methods can discover key repeating elements, such as jingles and separators, however they cannot infer the entire structure of a program. In this paper, we propose a hierarchical use of grammatical inference to yield a temporal grammar of a program from a collection of episodes, discovering both the vocabulary of the grammar and the temporal organization of the words from the vocabulary. Using a set of basic event detectors and simple filtering techniques to detect repeating elements of interest, a symbolic representation of each episode is derived based on minimal domain knowledge. Grammatical inference based on multiple sequence alignment is then used in a hierarchical manner to provide a temporal grammar of the program at various levels of details. Experimental validation is performed on 3 distinct types of programs on 4 datasets. Qualitative analyses show that the grammars inferred at the different levels of the hierarchy are relevant and can be obtained from a fairly limited number of episodes.

***Index Terms*—** TV program structuring, hierarchical structural grammar, symbolic representation, multiple sequence alignment, unsupervised and multimodal approach

1. INTRODUCTION

The last decade has seen a rapid increase in multimedia content and large scale audiovisual archives are available for users and content providers. For instance, the French National Audiovisual Institute, a repository of French radio and television audiovisual archives, has more than five million hours of radio and television programs stored. However, such collections are useless in practice if they are not described and indexed so as to know what they contain and to provide easy access to a particular portion of their content. Therefore, indexing of such large-scale archives is indispensable for browsing, sharing and Internet diffusion.

In particular, program structuring, which this paper focuses on, is a crucial step for high-quality indexing. Especially, TV program structuring consists in obtaining a temporal segmentation of programs into their basic constitutive elements. Each program has its own way of being organized by editors. For example, TV news programs usually start with a brief outline of the reports, followed by an alternation of anchorperson's announcement of the upcoming topics and news reports. Most news programs end with interview segments, weather forecasts or program trailers. Program structuring aims at detecting the existence and the temporal boundaries, i.e., the start and end frames, of such constitutive elements, designated as the structural element, of the program. A *structural element* hence refers to a video segment with a particular syntactic meaning.

We focus on recurrent TV programs where program structuring is realized by leveraging the repetitiveness to discover recurrent structural elements from which to infer a temporal grammatical structure, i.e., structural elements and their temporal ordering. A recurrent TV program is a program with multiple episodes which are periodically broadcasted (e.g., daily, weekly), such as News, entertainments and magazines. Most of their episodes follow the same editorial structure: Structural elements appear in almost the same order with very similar duration, separated by certain short sequences (named *separators*) which repeat across episodes at similar time instants. However, different types of TV programs lead to diversity of program structures, so identifying such structural elements and separators is not a trivial task. In other words, the challenge lies in inferring grammatical structure with minimal prior knowledge about the program genre and the types of structural elements which may be present. Facing the challenge, based on a collection of episodes from the program, we discover elements relevant to the structure by means of repeatability of recurrent programs and infer the corresponding grammar. Thus a model of the structure of the program can be established to process additional episodes.

To achieve this goal, adopting an unsupervised framework of TV program structuring as presented in [1], we propose a hierarchical architecture for grammatical inference based on multimodal structural elements. Specifically, the framework proposed in [1] consists of three steps, i.e., generic visual event detection, structural element identification, basic grammatical inference. In this paper, we first extend the generic event detection and structural element identification approaches from the sole visual modality to audiovisual modality in order to increase the completeness of the program structure. We then adopt a top-down architecture for grammatical inference to obtain hierarchical structural grammars at different granularity, and integrate a sectional-aligned method where multiple sequence alignment techniques is piece-wise applied on grammar inference to ameliorate the disambiguity of the inferred grammar. The sectional-aligned method is the method that multiple sequence alignment techniques is applied to each of the per-determined intervals of a program while inferring grammars. The top-down architecture allows to infer diverse structural grammars from coarse-grain to fine-grain, and the sectional-aligned method provides more deterministic grammars by aligning shorter sequences.

The paper is organized as follows. Section 2 reviews the exist-

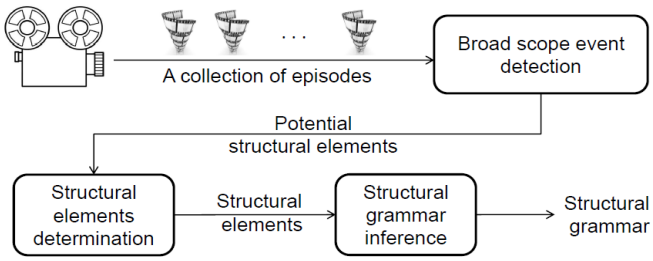


Fig. 1. General architecture of the three step approach for the grammatical inference of a program structure.

ing techniques for TV program structuring. Section 3 explains the overall method and details each step of structural grammar inference for unsupervised TV program structuring. Experimental evaluations are reported in Section 4, followed by conclusions in Section 5.

2. RELATED WORK

TV program structuring has been extensively studied. Existing approaches can be classified in two categories, according to whether prior knowledge of structural elements is incorporated in program structuring or not.

Most previous studies on TV program structuring focused on the case where information on the structure is available as prior knowledge. Resorting on prior knowledge of the structure, research has targeted either on entire structure of the program (e.g., [2], [3]) by leveraging structure models learned from annotated data, or on typical structure elements (e.g., [4], [5]) by considering their inherent properties across diverse types of programs. Hidden Markov models have been widely used in the first case [6], [7], while event detection has been applied to both cases. Most of these approaches are however supervised as they rely on the prior knowledge of the structure. They depend on genre of the program or type of the structural element.

As an alternative, another category of very recent approaches structures programs without prior knowledge of the structure. As not knowing the constituting structural elements, the structure of the program is discovered by means of event repetitiveness (e.g., visual recurrence [8] and audio recurrence [9]), sequential model [1], frequent pattern [10] or audiovisual consistency [11]. All these approaches are used for recovering the underlying structure of a program. The approaches of this category try to skirt supervised techniques and adopt minimal prior knowledge in order to preform unsupervised program structuring.

Work reported in this paper follows the last path and attempts at discovering the structural elements from a collection of episodes, along with a model for their sequential nature, thus targeting structure completeness and program type diversity.

3. HIERARCHICAL STRUCTURAL GRAMMAR INFERENCE

In this paper, we address the problem of inferring a hierarchical structural grammar for recurrent TV programs from a collection of episodes of the same program based on minimal prior knowledge. To avoid any confusion, we propose to use the term *program* to refer to a recurrent TV program and *episode* to refer to an exemplar of the program. A structural grammar consists of the set of structural elements composing each episode, and their temporal order. For example, reports and anchorperson's announcements are structural elements for a news program, on top of which a temporal model with occurrence probabilities can be built to form a structural grammar. To achieve the temporal model discovery, symbolic representation of structural elements is proposed for grammatical inference. Especially, in the absence of prior knowledge of which structural elements compose the program, we adopt the property of event recurrence across all episodes to determine the structural elements, represented as symbols, in an almost unsupervised manner. A structural grammar can then be inferred by multiple sequence alignment techniques.

We propose a three step approach as illustrated in Figure 1. Firstly, a set of broad scope events are detected within each episode by a number of audiovisual detectors. These events might be of interest as potential structural elements. In the second step, we assess the set of events detected along two lines. Role recognition is to identify the important persons of each episode, while density estimation is used to find events which recurrently occur at about the same instant in each episode. These two strategies are further used to assess repeatability of structural elements so as to help in deriving a symbolic representation by adopting minimal domain knowledge. Finally, we infer the hierarchical structural grammar of the program by leveraging multiple sequence alignment techniques, where a top-down and sectional-aligned architecture is adopted.

In the following subsections, we will detail each step of our approach.

3.1. Multimodal detection of broad scope events

In order to detect general purpose events which serve for discovering generic enough structural elements for different types of shows, a large number of event detectors should be adopted. Practically, considering a trade-off between the type of programs, the complexity at run time and implementation, a number of key detectors are applied.

Five visual detectors described in [1] were applied. Shot detector and dissolve detector are utilized to detect shots. A monochrome image detector is used for detecting the most evident separators. Person clustering classifies people appearing in each episode into different groups, and shot reverse shot detector is applied based on person clusters. Shot reverse shot depicts shots alternating between two characters facing one another, which are usually engaged in dialog scene or some face to face interaction.

In this paper, four detectors are added to improve the completeness of program structure, where text detector and motion activity detector are based on visual content, and speech/music/silence detector and audio recurrence detector are based on audio content. Text detector aims at detecting and localizing text(-like) regions in images. Motion activity estimation is often associated with text region localization to detect still text scene. We adopt a block matching algorithm [12] to estimate the motion activity in current frame with respect to the previous frame of a video. Speech/music/silence detector allows evidently to identify different types of audio sequences. Music might refer to songs, jingles, etc. Evident silences might appear in commercials segments, (before or after) separators, etc. Outlines for a program, i.e., a brief summary of the main points of a show, are mostly found when speech and music appear at the same time. Speech/music/silence detector is performed by the implementation described in [13]. Audio recurrence detector, based on the MODIS software [14], aims at detecting identical audio sequences, which allows to detect separators as they repeat within or among different episodes of a program.

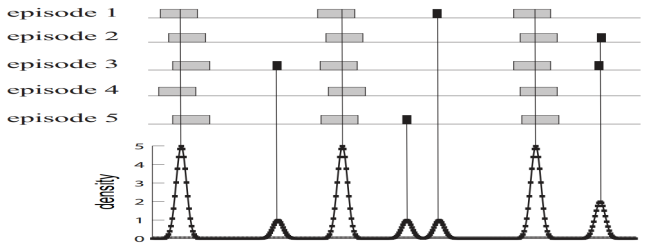


Fig. 2. Example of temporal density filtering to select structural events corresponding to structure separators (from [8])

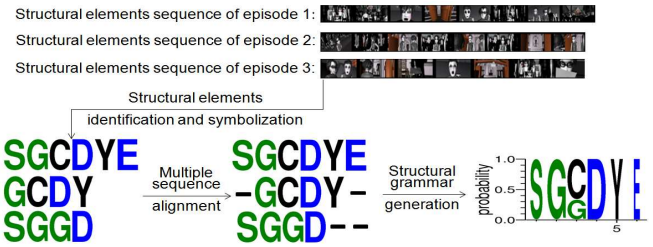


Fig. 3. Illustration of structural grammar inference from three episodes with symbols SGCDYE.

3.2. Determination of structural elements

A considerable amount of events are detected in the first step, but they are not all relevant to the structure of recurrent programs. For example, a sequence of monochrome images connecting two parts of a game show is a separator while it can also be found in a night scene. In order to remove the spurious structural elements, this step addresses the determination of valid structural elements for the structure of a program, from which a symbolic representation of episodes is obtained. Determination of structural elements implements two complementary strategies: role recognition and density filtering.

Role recognition is adopted to further characterize the outcome of person clustering and identify important persons. We mostly focus here on the conductor, or anchor, role which is clearly a strong cue with respect to structure. Five different features are defined to characterize each person cluster, viz.: total duration of appearance; total number of distinct appearances, i.e., number of non consecutive segments; duration of the longest segment in which the person appears; time range between the first and last occurrence; duration in which the speaker is engaged in a dialog. To account for varying episodes and program lengths, all five metrics are scaled to $[0, 1]$. Decision on the dominant person is made based on the sum of the five normalized metrics. The cluster for which the sum is maximal is identified as the dominant person, as a dominant person is assumed to ideally take significant time of an episode.

Following the role recognition, density filtering is to analyze the temporal distribution of the events which repeat with relative temporal stability across episodes, i.e., valid structural elements. For instance, repeated short audio sequences with similar temporal position across episodes are considered as separators, while a long shot containing dominant person is supposed to be anchor's opening if it is found at the beginning for most episodes of the program. Valid structural elements are identified by taking advantage of the repeatability for recurrent TV programs, as they share relatively stable temporal structures across different episodes which have almost identical structural elements. For a collection of episodes from a program, we project the occurrences of each kind of detected events onto the same temporal axis, filtering out events which occasionally appear in some episodes by counting the number of occurrences. Temporal density analysis with a kernel function as follows is adopted to select the valid structural elements:

$$f(x;h) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \quad (1)$$

where $f(x;h)$ is the estimated density function, x is the data from which the density estimate is constructed, h is the bandwidth and $K()$ is the kernel. A Gaussian kernel is used with optimal bandwidth automatically chosen [15]. An illustration for density filtering is given in Figure 2, where events in black are removed while the ones in gray are considered as structural elements because of their high temporal concentration. The property of repetitiveness of structural elements across episodes gives the explanation for structural elements selection based on thresholding of density function.

3.3. Top-down sectional-aligned structural grammar inference

Recurrent events determined in the previous step, are identified as structural elements by adopting minimal prior knowledge of TV programs. For instance, a structural element corresponding to a sequence of white frames is a separator, while a long duration shot containing the dominant person at the beginning of a program is the conductor's opening. As a result of structural elements determination, symbolic representation is derived. Each element is represented as an alphabet symbols, while each episode of a program can then be represented as a time-ordered sequence of symbols. The temporal stability of recurrent TV program structure results in similar symbolic sequences across episodes. However, slight differences still exist among different episodes. In order to infer a structural grammar for the program from such sequences, we adopt multiple sequence alignment techniques to discover the common pattern across symbolic sequences. Multiple sequence alignment techniques allow to align the symbolic sequences in the way that alphabet symbols, i.e., valid structural elements, in a given position are homologous, superposable or play a common functional role. We adopt the ClustalW algorithm explained in [16] as multiple sequence alignment tool. The process of grammar inference from a symbolic representation is illustrated in Figure 3 with three episodes. Symbols, i.e., $SGCDYE$, represent the valid structural elements that are identified. we mark that the alphabet letter for each structural element is chosen manually. A graphical representation of the resulting grammar is obtained using WebLogo [17]. A stack of symbols is used to illustrate each position in the grammar: The height of objects within the stack indicates the relative frequency of each symbol while the stack width is proportional to the fraction of valid symbols in that position.

Comparing with the induced structural grammars in [1] which are relatively concise, more complete structural grammars are in this paper inferred, since multimodal event detectors are applied to programs. Increased structural completeness and complexity leads evidently to more ambiguities for structural grammars in the process of multiple sequence alignment. To reduce the grammar ambiguity and infer grammars at different granularity, we propose hierarchical structural grammar inference with a top-down and sectional-aligned architecture. Firstly, separators are determined by filtering recurrent events whose durations are significantly short as well as adopting prior knowledge of most common separators, such as monochrome images and repeated audiovisual sequence. Based on separator determination, a coarse-grain grammar is obtained with only separators and chapters, i.e., video segments between two separators, whose

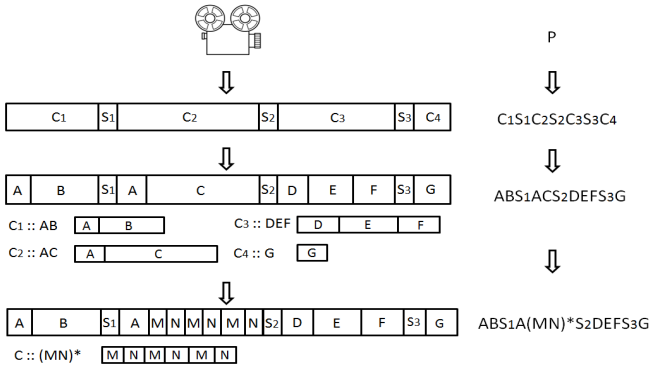


Fig. 4. A top-down and sectional-aligned architecture for hierarchical structural grammar inference

structural grammar is to be discovered. Secondly, the structural grammar of each chapter is inferred by leveraging the basic grammar inference explained in the previous paragraph, and a fine-grain structural grammar can be obtained. Moreover, we can still infer a grammar for each detected structural element if it can be decomposed into parts of different elements. Thus a finer-grain structural grammar can be obtained. The top-down and sectional-aligned architecture is illustrated in Figure 4. A program is firstly decomposed only into separators represented by S_i and chapters represented by C_i. The grammars of each chapters are then inferred: The grammar of chapter C_1 is composed of two structural elements A and B; the grammars of C_2, C_3 and C_4 are also obtained. The structural elements are represented by different alphabet letters, i.e., $ABCD$..., where P stands for the entire program. At last, we can go deeper to discover the grammars of structural elements if any. For example, structural element C can be decomposed into alternation of elements M and N, therefore a detailed grammar can be inferred.

In short, a program is decomposed from coarse elements to greater details, i.e., top-down approach, and during each step of decomposition, grammars of different levels of elements are inferred, i.e., sectional-aligned approach. The sectional-aligned grammar inference replaces aligning long sequences of entire programs by aligning short sequences of certain segments of a program, which obtains a more deterministic structural grammars. Besides, grammars at different granularity from coarse grain to fine grain can be obtained at the same time.

4. EXPERIMENTAL RESULTS

Experiments are conducted on four recurrent programs from three different types, viz., game, news and magazine. We firstly demonstrate structural element determination. Then the qualitative analysis of inferred grammar is performed by varying the number of episodes used for grammar inference.

4.1. Dataset description

Four different programs, with 24 episodes each, are used for inference and evaluation, as given in Table 1. *20h News* (NEWS), a daily news show of 2007, follows a very standard pattern for such programs. Two programs of type magazine, *Telematin* (MAGM) and *Thalassa* (MAGS), were taken with episodes selected from 1989 and 1997. MAGM is a morning program proposing news and topics about culture and daily life. Different topics of the program are separated by separators. MAGS is a magazine about sea stories, where a conductor leads the show which is composed of reports and discussions. *Que le meilleur gagne* (GAME) is a game show whose episodes were taken over two years (1991 and 1992). It has four parts divided by separators. The program, hosted by a conductor, mainly contains interview scenes and question/answer scenes with full text segments. While the same conductor appears in all the episode of GAME, more than two distinct conductors can be found for NEWS, MAGM and MAGS.

Table 1. Description of the datasets used for evaluation

Dataset	Date	Episodes	Type	Average duration
GAME	1991 - 1992	24	Game	31.9 m
NEWS	2007	24	TV news	37.9 m
MAGS	1997	24	Magazine	57.7 m
MAGM	1989	24	Magazine	61.9 m

4.2. Performance of structural elements determination

A number of structural elements were determined from a scope of recurrent events with minimal prior knowledge. Different types of structural elements are presented in Table 2, where the corresponding event detectors of each structural element are listed. Specifically, if an event detector contribute to identifying a specific structure element, it is marked as "yes" or with its corresponding detected results. As described in previous sections, density filtering is applied to find the recurrent elements across episodes. Some structural elements are determined by a sole event detector such as dialogs, music/songs or separators for NEWS, MAGS and MAGM. Some structural elements are discovered by more than two event detectors. In other words, the high temporal concentration of density functions of events are coincident. Person cluster results with long shot duration are used to find person's monologues, anchor person's monologues can also be found if we count in dominant person prediction. Besides, commercials are characterized by the presence of (any combination of) black frame, short shot duration and audio silence. In NEWS the outlines pronouncing the main topics in each episode always has short shot duration and anchor person's speech with back ground music. Effectiveness of structural element determination is detailed in [1], where a high detection recall for structural elements are presented.

4.3. Hierarchical structural grammar inference

The inferred grammars are illustrated in Figure 5 for the four programs. Separators are firstly determined, and the coarse-grain grammars with separators and chapters are obtained. Chapter is represented as N, whose structural grammar is to be inferred. We denote separator as S for NEWS, MAGS and GAME since each of them have identical separators, while we denote different separators of MAGM as S,V,F,H,G,I,K because of diversity of its separators: The separators are found in different clusters of repeated audio sequences. We mark that although separators for both MAGS and MAGM are detected by audio recurrence, the difference is that the separators of MAGS belong to the same recurrent audio class while the separators of MAGM are found in different recurrent audio classes. For further discovery, based on the hierarchical method described in section 3.3, the chapters are then structured and finer-grain grammars are inferred. For NEWS, illustrated in Figure 5(a), after determining the separators, we then determine the outlines. The segment after outlines, accounting for most time of the program, is

Table 2. Structural elements detection using a broad scope of detectors

Structural element	Symbol	Shot duration	Dissolve transition	Mono-chrome image	Centra-lized text	Motion activity	(Dominant) person	Shot reverse shot	Speech music silence	Audio recurrence
Separator GAME	S		yes	yes						
Separator MAGM	S,V,F,H,G,I,K									yes
Separator MAGS	S									yes
Separator NEWS	S			yes						
Dialog	D							yes		
(Anchor's) monologue	(A) L	long					yes			
Music/song	M								music	
Commercials	P	short		yes					silence	
Outline NEWS	T	short							music&speech	
Full screen text	E				yes	low				

considered as news content. Based on anchorperson prediction, news content can then be divided into alternating patterns and anchorperson's introduction and news report, resp. denoted as A and R. As the number of reports varies across news episodes, we use just one occurrence to represent the repeated alternation. For GAME in Figure 5(b), besides separators, dominant person (i.e., monologues of the conductor) and dialogs are the valid structural elements, resp. denoted as A and D. The segments in the second level of the grammar, denoted as N, can be further determined for different episodes, as dialogs or/and full screen text denoted as E. From the grammars, the main syntax of GAME can be explained as: the game starts with an introduction (separator) followed by a dialog (between the anchor and the participants). We then have an alternation of person interacting (dialogs) and game phases (full screen texts), or one of them. For MAGS, illustrated in 5(c), selected symbols are based on dominant person's monologue and dialog segments, yielding a relatively deterministic grammar. A continuous segment with long duration, denoted as N, is considered as a report while A and D represent anchorperson's monologue and dialog respectively. For MAGM, the grammars are in Figure 5(d), where anchorperson's opening, music, dialog, monologue, full screen text and commercials are determined structural elements, resp. denoted as A, M, D, L, E and P. MAGM is divided into many chapters by separators, and content of each chapter varies a lot. The inferred grammars of different granularity represent the structures of different levels of details.

In order to analyze the stability of grammars inferred, i.e., quantity of determined structural elements and their temporal stability, we vary the number of episodes used to infer the grammars. 6, 12, 18 and 24 episodes are chosen respectively to infer grammars for each program and the results are illustrated in Figure 6. We observe that with less episodes we could have more separators detected. Typically, in Figure 6(d), the grammar of MAGM inferred by 6 episodes has 8 separators, while 6 for 12 episodes and 5 for 18 and 24 episodes. This can be explained by the fact that separators are very short audio or visual sequences easily drowned in a large number of episodes. The same phenomenon can be found in grammars of GAME in Figure 6(b), the proportion of separators reduces when the number of episodes increases. This result shows that a small collection of episodes usually results in a high recall of separator detection. On the contrary, the numbers of detected structural elements tend to be stable along with the augmentation of the number of episodes, and the grammars converge, as shown in Figure 6(a), 6(b), 6(c). For MAGM in Figure 6(d), grammar stability is less evident, for the structure is more complex than others. However, the grammars stay relatively stable for those two inferred by 18 and 24 episodes. In order to infer a complete grammar which is generic enough to represent the structure of the program, these two phenomena hint that we may vary the number of episodes used to infer the grammar according to the wanted granularity. A small quantity of episodes, like 6 episodes, is sufficient for a coarse-grain grammar with only separators and chapters. For inferring a finer-grain grammar, the appropriate quantity of episodes is the number for which the grammar of program tend to converge. For instance, 6 episodes for GAME, 12 episodes for both NEWS and MAGS and 18 episodes for MAGM.

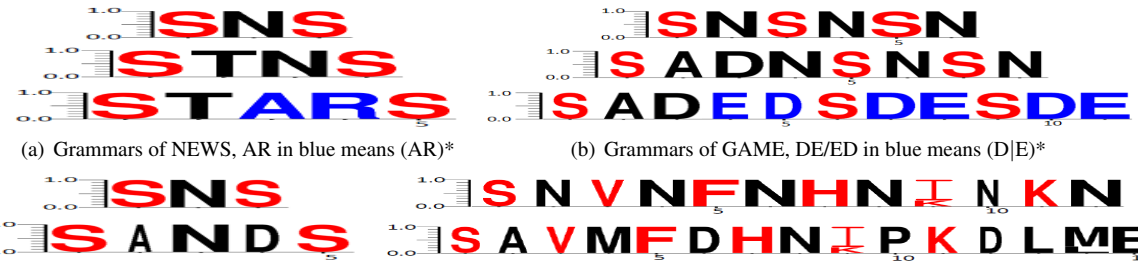
5. CONCLUSION

The structure discovery work described in this paper shows that a hierarchical structural grammar inference can efficiently yield an overview model for program structure. Moreover, the symbolic representation can serve for grammar inference for recurrent programs from a collection of episodes, with almost no supervision and no specific training data. In this work, a broad scope of event detectors of multimodality are firstly applied, followed by structural elements determination with minimal prior knowledge by exploiting role recognition and recurrence across episodes. A top-down and sectional-aligned architecture is then designed for grammatical inference to obtain hierarchical structural grammar at different granularity with minimal ambiguities. The event detection and selection techniques used are generic enough to be useful for a large variety of programs. The generalization of our approach lies in automatic discovery of the structure of recurrent TV programs. Experimental evaluation on four datasets of three types of programs shows that the grammar of program brings a final layer of structure of the program. By varying the number of episodes used for grammatical inference, a more relevant structure can be obtained for different granularity. Once the grammars of a program is obtained, the grammatical model can be utilized to structure additional episodes from the same program.

6. REFERENCES

[1] Bingqing Qu, Félicien Vallet, Jean Carrive, and Guillaume Gravier, "Using grammar induction to discover the structure of recurrent tv programs," in *International Conferences on Advances in Multimedia*, 2014.

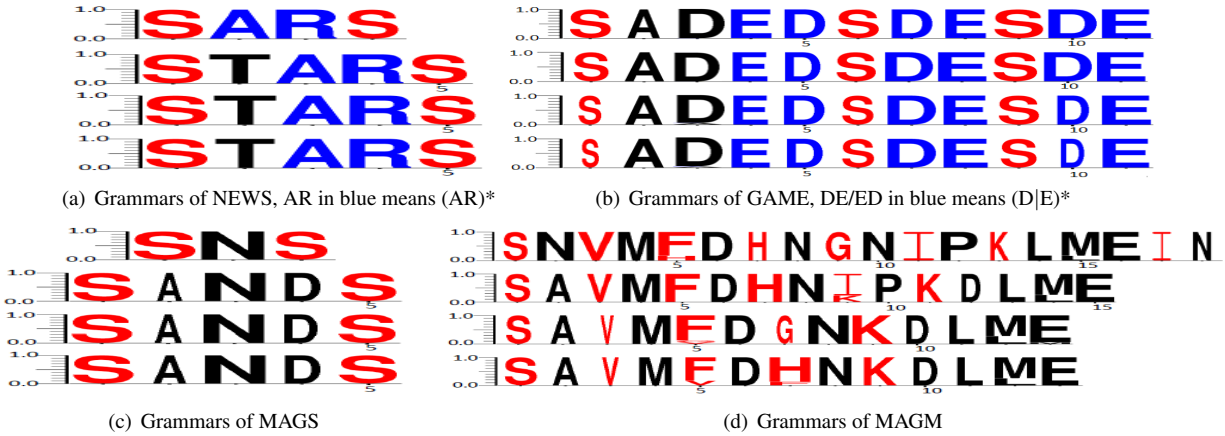
[2] Marco Bertini, Alberto Del Bimbo, and Pietro Pala, "Content-based indexing and retrieval of tv news," *Pattern Recognition Letters*, 2001.



(a) Grammars of NEWS, AR in blue means (AR)* (b) Grammars of GAME, DE/ED in blue means (D|E)*

(c) Grammars of MAGS (d) Grammars of MAGM

Fig. 5. Grammars inferred using 12 episodes for resp. NEWS, GAME, MAGS and MAGM, each of which are coarse-grained to fine-grained grammar from top to down.



(a) Grammars of NEWS, AR in blue means (AR)* (b) Grammars of GAME, DE/ED in blue means (D|E)*

(c) Grammars of MAGS (d) Grammars of MAGM

Fig. 6. Grammars for resp. NEWS, GAME, MAGS and MAGM. From top to down of each, the grammars are inferred using resp. 6, 12, 18 and 24 episodes.

[3] Haojie Li, Jinhui Tang, Si Wu, Yongdong Zhang, and Shouxun Lin, "Automatic detection and analysis of player action in moving background sports video sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, 2010.

[4] Alan Hanjalic, R.L. Lagendijk, and Jan Biemond, "Template-based detection of anchorperson shots in news programs," in *International Conference on Image Processing*. IEEE, 1998.

[5] Xinbo Gao and Xiaoou Tang, "Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing," *IEEE Transactions on Circuits and Systems for Video Technology*, 2002.

[6] Lexing Xie, Peng Xu, Shih-Fu Chang, Ajay Divakaran, and Huifang Sun, "Structure analysis of soccer video with domain knowledge and hidden markov models," *Pattern Recognition Letters*, 2004.

[7] Ewa Kijak, Guillaume Gravier, Lionel Oisel, and Patrick Gros, "Audiovisual integration for tennis broadcast structuring," *Multimedia Tools and Applications*, 2006.

[8] Alina Elma Abduraman, Sid-Ahmed Berrani, and Bernard Merialdo, "An unsupervised approach for recurrent tv program structuring," in *European Interactive TV Conference*, 2011.

[9] Armando Muscariello, Guillaume Gravier, and Frédéric Bimbot, "Audio keyword extraction by unsupervised word discovery," in *INTERSPEECH 2009: 10th Annual Conference of the International Speech Communication Association*, 2009.

[10] Arne Jacobs, "Using self-similarity matrices for structure mining on news video," in *Advances in Artificial Intelligence*. Springer, 2006.

[11] Mathieu Ben and Guillaume Gravier, "Unsupervised mining of audiovisually consistent segments in videos with application to structure analysis," in *IEEE International Conference on Multimedia and Expo*, 2011.

[12] Aroh Barjatya, "Block matching algorithms for motion estimation," *IEEE Transactions Evolution Computation*, 2004.

[13] Costas Panagiotakis and Georgios Tziritas, "A speech/music discriminator based on rms and zero-crossings," *IEEE Transactions on Multimedia*, 2005.

[14] Laurence Catanese, Nathan Souviraà-Labastie, Bingqing Qu, Sebastien Campion, Guillaume Gravier, Emmanuel Vincent, and Frédéric Bimbot, "Modis: an audio motif discovery software," in *Show & Tell-Interspeech*, 2013.

[15] Zdravko Botev, Joseph Grotowski, and DP Kroese, "Kernel density estimation via diffusion," *The Annals of Statistics*, 2010.

[16] MA Larkin, Gordon Blackshields, N.P. Brown, R Chenna, Paul A McGettigan, and Hamish McWilliam, "Clustal w and clustal x version 2.0," *Bioinformatics*, 2007.

[17] Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner, "Weblogo: a sequence logo generator," *Genome research*, 2004.